LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG

Manual
CisWeb application
A gwt-based web service that provides access to language processing tools of CIS

Authors:
Version 1.0 Luitpold Gollas
Version 2.0 Sebastian Bildner
Version 3.0 Patrick Seebauer

APPs from Authors: Eduardo Torres, Jörg Schuster, Sebastian Nagel, Max Hadersbeck, Thorsten Voblt, Florian Fink

Project Manager: Dr. Max Hadersbeck

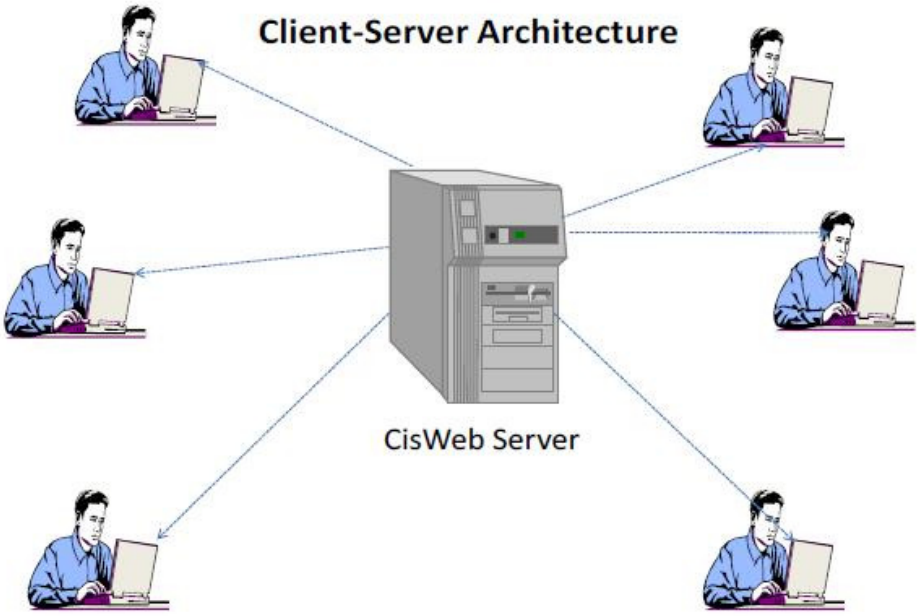
1. Introduction

CisWeb is a web service that provides access to several language processing tools developed at the “Centrum für Informations- und Sprachverarbeitung München” (CIS). The development was started during an internship of Luitpold Gollas at the professorship for Programming and Software techniques (PST) at the LMU from March 07 to August 07. The corner stone of the web service was laid by an application, developed at the CIS, called CisTex, a standalone JAVA-application which has to be installed from the user on his client and runs locally. That means that every user had to download the whole suite before he could start working with it and every executable had to be compiled for his operating system.

The next step in developing a workbench for computer linguists was to offer a web service that holds the linguist tools on a server and the user communicates with the tools via an internet browser. Everybody should have access to the tools, without downloading the whole suite. (see Picture 1).

Advantages of this approach are that the tools have to be updated only on the server and every user has access to the latest versions all the time. Furthermore for the CIS it is sufficient to compile only a Linux version of the executables. The users, who work with the web service, are only provided with a graphical interface that will be shown in their browser window. With this graphical interface, the calls to the server side tools are performed and the server’s results are presented as graphical output within the user’s browsers or stored as files in a personal home directory on the server. To maintain a home directory for every user, the users must register them self if they start the web service at the server.

Client-Server Architecture



Picture 1: Client-Server Architecture of CisWeb

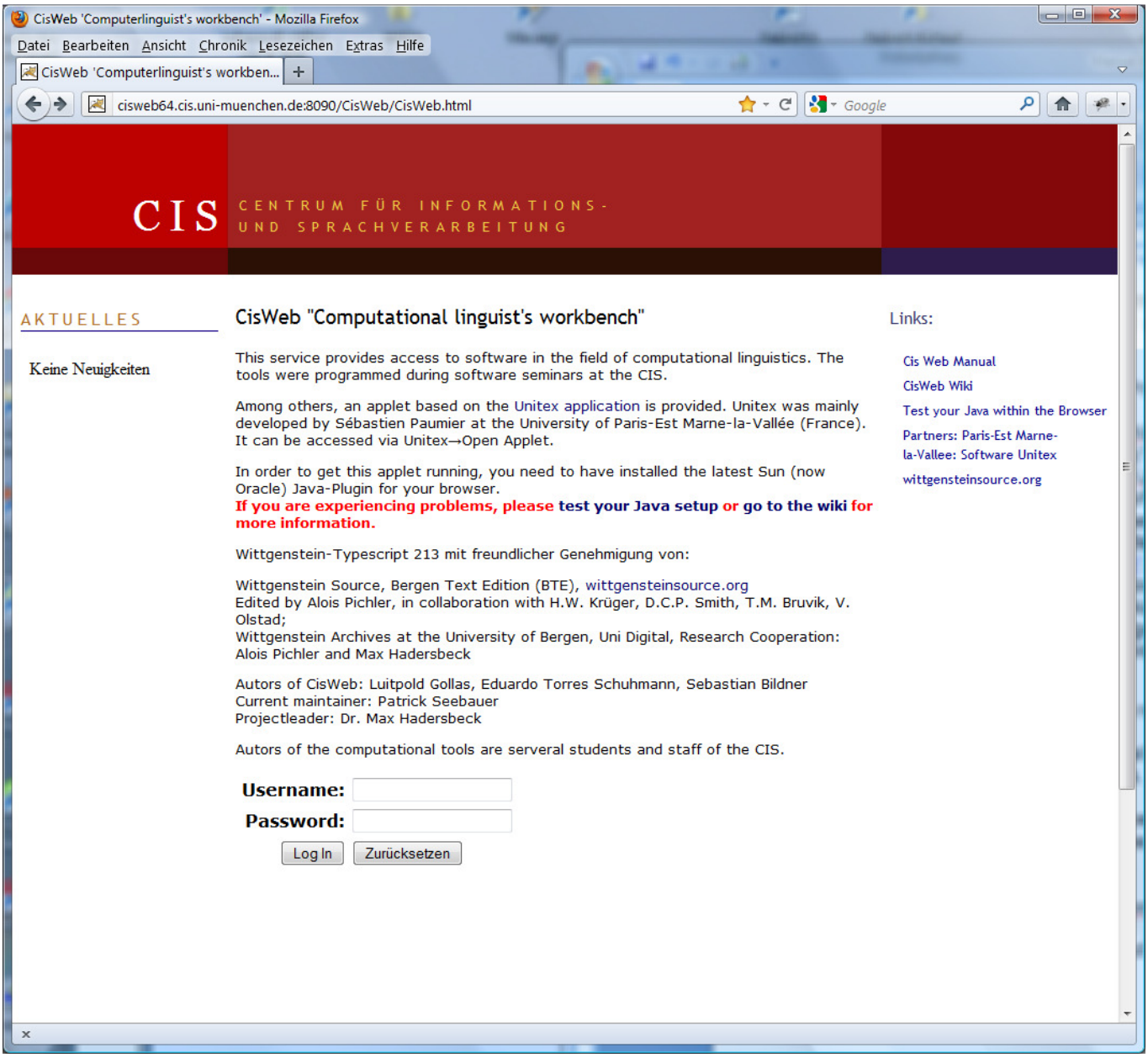
As CisWeb is a Client-Server application it has to satisfy more demands than a local application like CisTex or Unitex. CisWeb offers user management for different roles like administrators, tutors or students. Every registered user is associated to his own folders where he can upload, download, compress, uncompress, view or delete files. Beside the private files, the web service offers a public area, where files can be exchanged between the administrator and the users. With the help of the CisWebinterface, the users can start end-of-sentence-recognition, concordance-calls and language-recognition on these files and the results can be seen in the browser window. As a scenario, one could imagine that a tutor provides some text files in the public folder of the web service so that every student in class has access to these files. He can work on the files and upload the results again. Due to the role management the tutor can collect the results from the student's folders. A short description of how one can use the CisWeb will be provided in the following sections.

The workbench is based on the public Domain Tools:

gwt-Google Web Toolkit: http://code.google.com/intl/de-DE/webtoolkit/ .
Apache Derby Database: http://db.apache.org/derby/ and
Apache Tomcat http://tomcat.apache.org/

2. Login Screen

The entry point address is http://cisweb.cis.uni-muenchen.de. This lead so a forwarding to the real address of CisWeb, which may be changing sometimes. The first screen, you will see, will be a login screen (Picture 2).

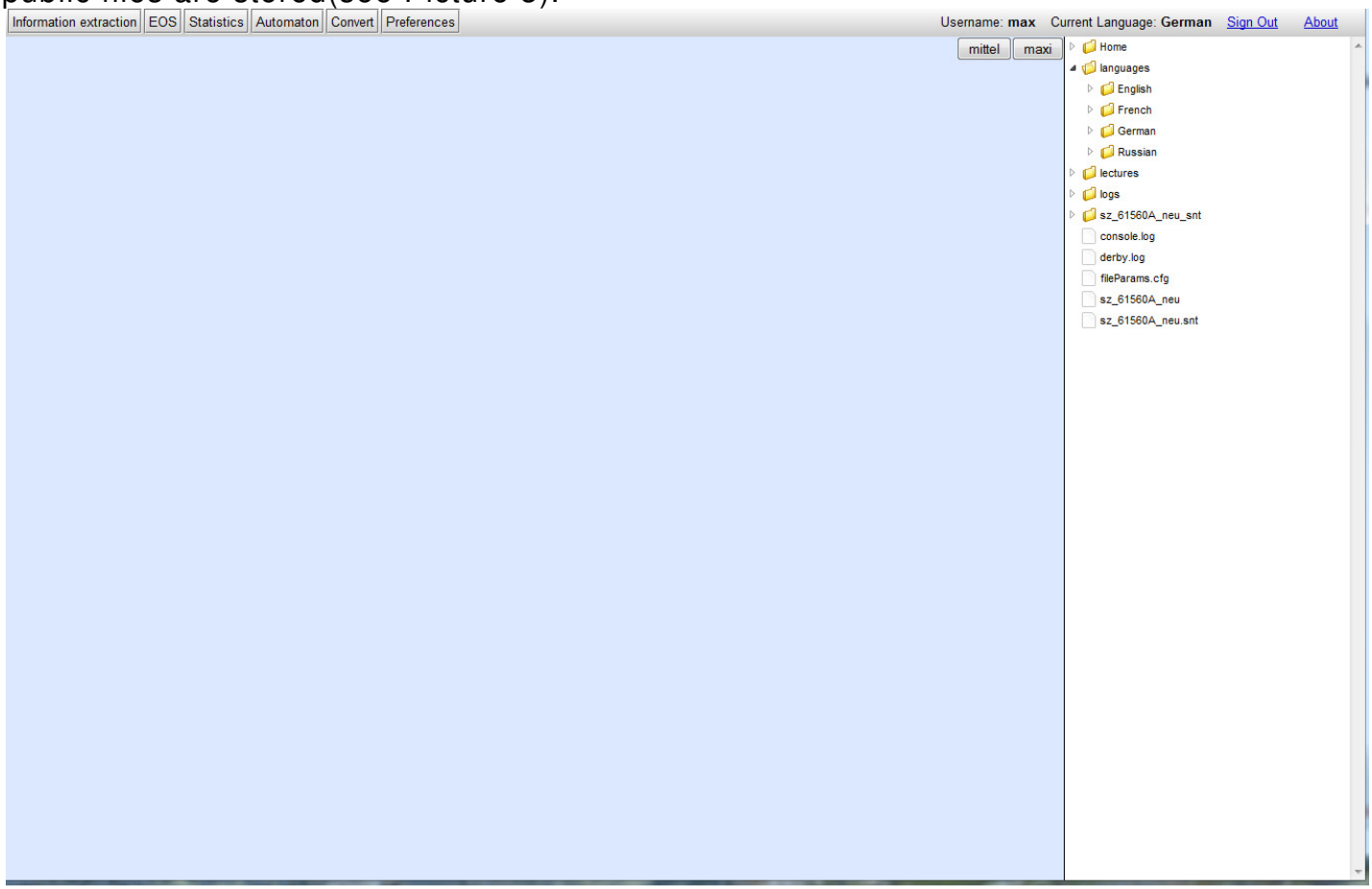

CisWeb 'Computerlinguist's workbench' - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
CisWeb 'Computerlinguist's workben...
cisweb64.cis.uni-muenchen.de:8090/CisWeb/CisWeb.html
Google
CIS
CENTRUM FÜR INFORMATIONS- UND SPRACHVERARBEITUNG
AKTUELLES
Keine Neuigkeiten
CisWeb "Computational linguist's workbench"
This service provides access to software in the field of computational linguistics. The tools were programmed during software seminars at the CIS.
Among others, an applet based on the Unitex application is provided. Unitex was mainly developed by Sébastien Paumier at the University of Paris-Est Marne-la-Vallée (France). It can be accessed via Unitex→Open Applet.
In order to get this applet running, you need to have installed the latest Sun (now Oracle) Java-Plugin for your browser.
If you are experiencing problems, please test your Java setup or go to the wiki for more information.
Wittgenstein-Typescript 213 mit freundlicher Genehmigung von:
Wittgenstein Source, Bergen Text Edition (BTE), wittgensteinsource.org
Edited by Alois Pichler, in collaboration with H.W. Krüger, D.C.P. Smith, T.M. Bruvik, V. Olstad;
Wittgenstein Archives at the University of Bergen, Uni Digital, Research Cooperation: Alois Pichler and Max Hadersbeck
Autors of CisWeb: Luitpold Gollas, Eduardo Torres Schuhmann, Sebastian Bildner
Current maintainer: Patrick Seebauer
Projectleader: Dr. Max Hadersbeck
Autors of the computational tools are serveral students and staff of the CIS.
Username:
Password:
Log In
Zurücksetzen
Links:
Cis Web Manual
CisWeb Wiki
Test your Java within the Browser
Partners: Paris-Est Marne-la-Vallee: Software Unitex
wittgensteinsource.org

Picture 2: Login screen of CisWeb

On this page, you can login to CisWeb and can test, if your Browser has installed JAVA-Plugins properly, because some tools use them for using complex graphical Applications (see : “In order to get this applet running, you need to have installed the latest Java-Plugin for your browser. If you are experiencing problems, please go to this site for testing your setup and for more information.”) To access CisWeb, you need a username and a password. To get your authentication data, ask the CIS-system administrator for a new account.
Click the “Log In” button to be forwarded to the CisWeb application. Authentication is required to establish a working directory on the CIS server for every user in order to save the private files you were working on that you can access them later again.

3. Main Screen

Until now, the main screen dispays the user interface of CisWeb that you will see all the time. All other features are shown in popup windows or dialog boxes. On the right side, you will see the content of your working directory. With the help of a Filebrowser you can access the files and navigate through your directories. Together with your working directory you have access to a public directory, were additional public files are stored(see Picture 3).



Picture 3: Main menu of CisWeb with Filebrowser and Menu

Furthermore you can upload files from your local client to your working directory on the server, with the “Upload file” box. If you log on the web service for the first time, your working directory may be empty because you have not uploaded files yet and there are no files in the public folder. You must click with the right button of your mouse, which has to be positioned over a Directory:

Picture 4: Upload File and Manage Directories

Left to the CisWeb Filebrowser you see the CisWeb Menu, with all of the CIS-Tools, that can be accessed over the web service. I will go more into detail on these tools in Features .

 Information extraction

This tool offers various tools to extract Phrases from a text file with the help of “Local Grammars”

a) Open Applet: Here you start the UNITEX Programingsystem, developed at: “University of Paris-Est Marne-la-Vallée (France)”, see: http://www-igm.univ-mlv.fr/~unitex/
b) Run Script: Here you can start the UNITEX Programingsystem from a PERL Script
c) Graph Editor … in work : This will offer soon a Graphical Editor for ‘Local Grammars’
d) Concordance: This implements a search (flotex) and build concordance tool
e) Suche: This search tool is connected to the Big Typescript 213 of Ludwig Wittgenstein, http://www.wittgensteinsource.org/ : “Wittgenstein-Typescript 213 mit freundlicher Genehmigung von: Wittgenstein Source, Bergen Text Edition (BTE), wittgensteinsource.org, edited by Alois Pichler, in collaboration with H.W. Krüger, D.C.P. Smith, T.M. Bruvik, V. Olstad; Wittgenstein Archives at the University of Bergen, Uni Digital, Research Cooperation: Alois Pichler and Max Hadersbeck”
f) EOS

This tool offers various end of sentence recognizer.

 Statistics

Here, you will find a tool that automatically detects the language of a text file and another tool, which can produce Frequency Lists.

 Automaton

A tool, that implements a Tgc (Transducer Grammar Compiler). It compiles Transducer-Grammatics to openfst-Transducers. The tools use the Opensource Transducerimplemenation from AT&T.

 Convert

Here you can Recode the Coding of a textfile and Strip HTML-Code from an .html File.

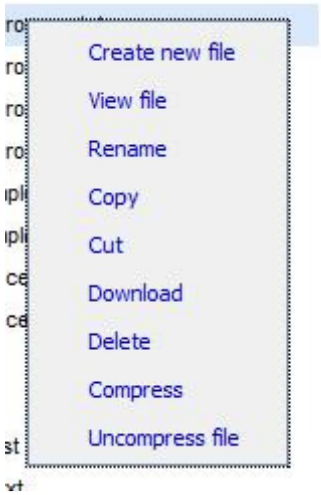
 Preferences

Here you can change your Password and select your Language, which is used from Unitex.

4. Features of CisWeb

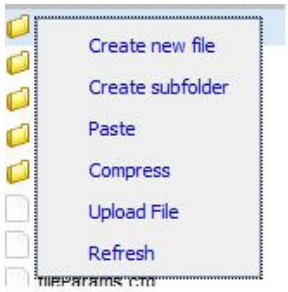
In the last section, I have already introduced all of the tools that are or will be integrated into CisWeb. When you read this manual, there might be other tools available because the application is still under development. Additionally to these tools, other features are implemented as well:

- File-operations



Picture 4: CisWeb FileMenu

When you click with your right mouse button on a fileicon within the Filebrowser, a small panel will appear that allows you to execute some operations on the file. Double Click on a File will view the content of the file. You will only see one part of the file because the server divides it into many parts and sends the parts to the client dynamically to improve the responsiveness. “Download file” is only a download link to the file.
“Delete file” deletes the file on the server, “Compress/Uncompress file” compress or uncompress the file (to .zip). Copy/Cut copy/pastes the selected File into a Clipboard and this File can be coped/moved to a new Directory, by clicking the right mouse button, situated over a directoryicon:



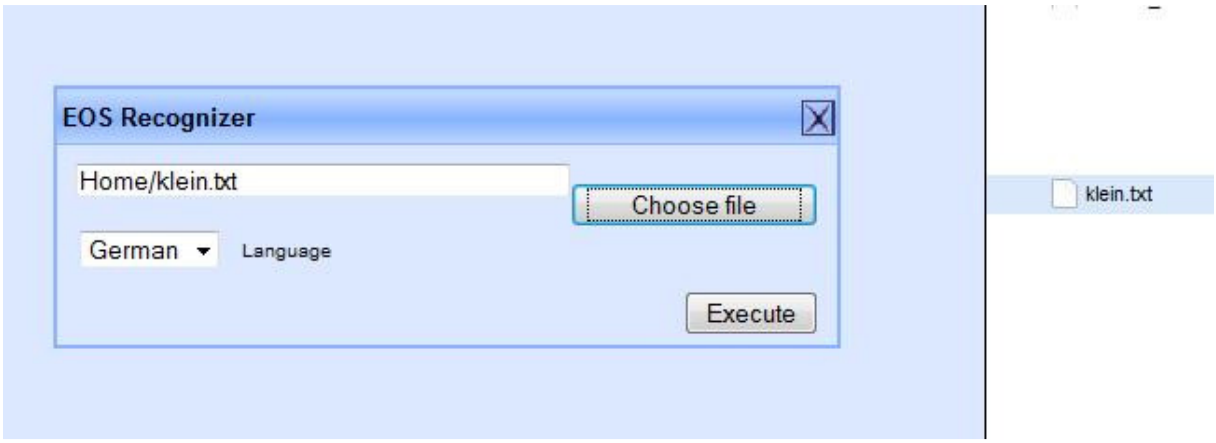
Picture 5: Paste a file into a new Directory

The user can also upload new files to the server. This can be done very easily by using the Upload File Option.

5. Computerlinguistic tools

5.1. Open files for the use within Tools

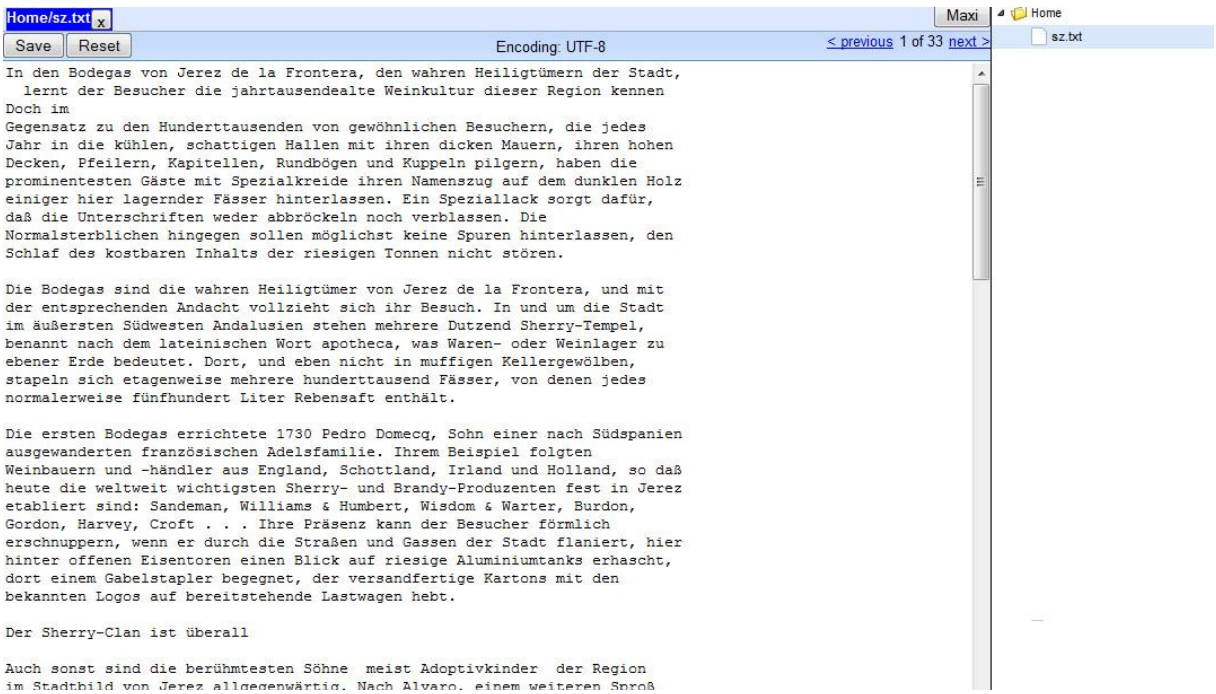
Furthermore the user can execute tools which work on his personal files, stored in his home directory.
Every tool works with Input-Files, which can be selected within Tool-Dialog-Box: You have to select a file in the fileview menu and click on “Choose file” in the Tool-Dialog-Box:



Picture 6: Choose Files for tools

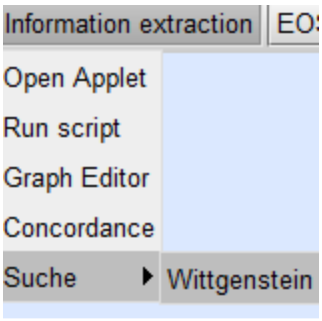
5.2. File Viewer

The file viewer (Picture 7: CisWeb File Viewer) can handle files of different encodings. The server automatically detects the encoding of the file and sends it with UTF-8 encoding to the client but the real encoding will be shown to the user. So, no matter in which encoding the file is stored at the server, the client always sees the right umlauts. As text files can be large and the downloading and displaying of the whole file would take a lot of time, the files are loaded dynamically. That means that only one part is downloaded to the client when he wants to view a file. The user navigates through the file by clicking on the “next”- or “previous”-link.


Home/sz.txt
Maxi
Home
sz.txt
Save
Reset
Encoding: UTF-8
< previous 1 of 33 next >
In den Bodegas von Jerez de la Frontera, den wahren Heiligtümern der Stadt,
lernt der Besucher die jahrtausendealte Weinkultur dieser Region kennen
Doch im
Gegensatz zu den Hunderttausenden von gewöhnlichen Besuchern, die jedes
Jahr in die kühlen, schattigen Hallen mit ihren dicken Mauern, ihren hohen
Decken, Pfeilern, Kapitellen, Rundbögen und Kuppeln pilgern, haben die
prominentesten Gäste mit Spezialkreide ihren Namenszug auf dem dunklen Holz
einiger hier lagernder Fässer hinterlassen. Ein Speziallack sorgt dafür,
daß die Unterschriften weder abbröckeln noch verblassen. Die
Normalsterblichen hingegen sollen möglichst keine Spuren hinterlassen, den
Schlaf des kostbaren Inhalts der riesigen Tonnen nicht stören.
Die Bodegas sind die wahren Heiligtümer von Jerez de la Frontera, und mit
der entsprechenden Andacht vollzieht sich ihr Besuch. In und um die Stadt
im äußersten Südwesten Andalusien stehen mehrere Dutzend Sherry-Tempel,
benannt nach dem lateinischen Wort apotheca, was Waren- oder Weinlager zu
ebener Erde bedeutet. Dort, und eben nicht in muffigen Kellergewölben,
stapeln sich etagenweise mehrere hunderttausend Fässer, von denen jedes
normalerweise fünfhundert Liter Rebensaft enthält.
Die ersten Bodegas errichtete 1730 Pedro Domecq, Sohn einer nach Südspanien
ausgewanderten französischen Adelsfamilie. Ihrem Beispiel folgten
Weinbauern und -händler aus England, Schottland, Irland und Holland, so daß
heute die weltweit wichtigsten Sherry- und Brandy-Produzenten fest in Jerez
etabliert sind: Sandeman, Williams & Humbert, Wisdom & Warter, Burdon,
Gordon, Harvey, Croft . . . Ihre Präsenz kann der Besucher förmlich
erschnuppern, wenn er durch die Straßen und Gassen der Stadt flaniert, hier
hinter offenen Eisentoren einen Blick auf riesige Aluminiumtanks erhascht,
dort einem Gabelstapler begegnet, der versandfertige Kartons mit den
bekannten Logos auf bereitstehende Lastwagen hebt.
Der Sherry-Clan ist überall
Auch sonst sind die berühmtesten Söhne meist Adoptivkinder der Region

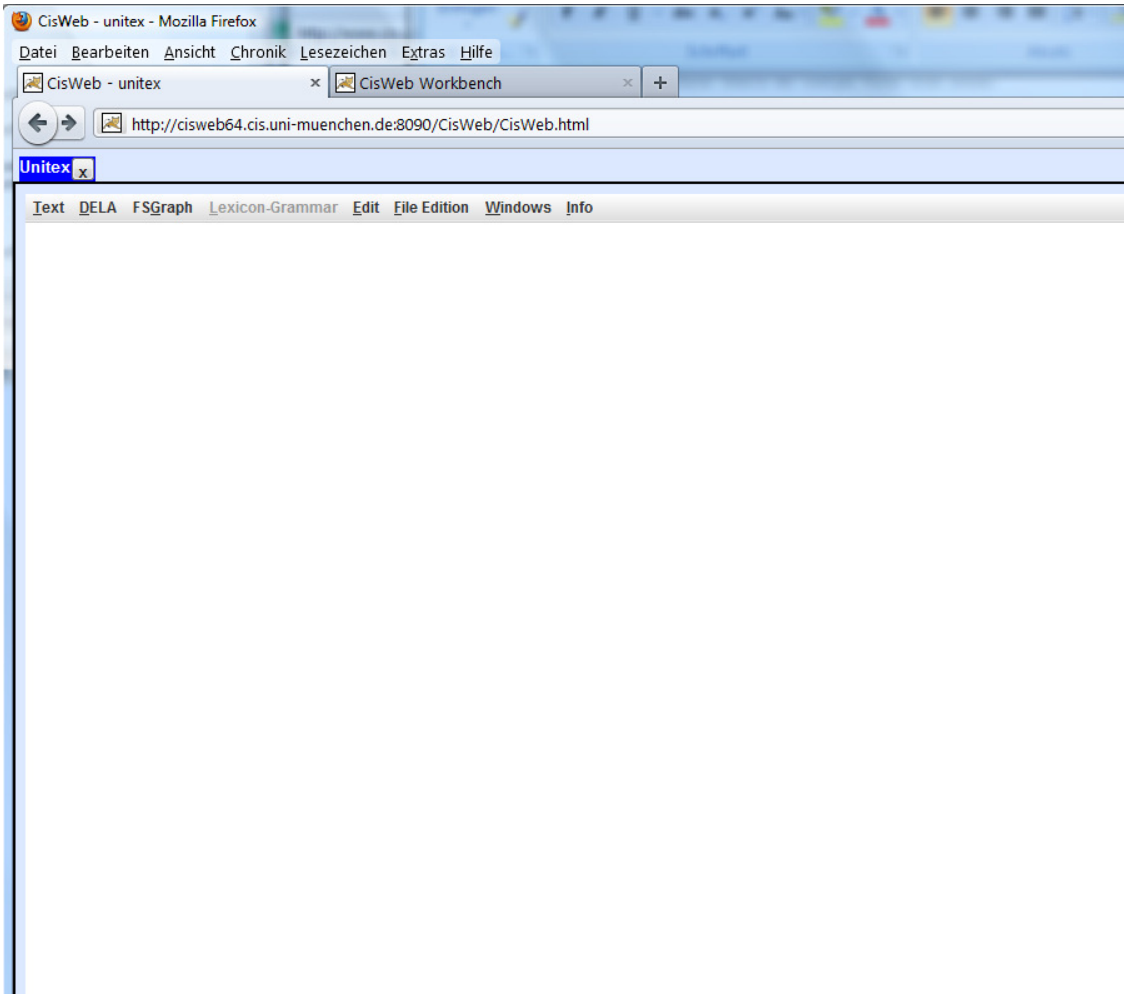
Picture 7: CisWeb File Viewer

5.3 Information extraction



Menu allows you to start 4 different tools:

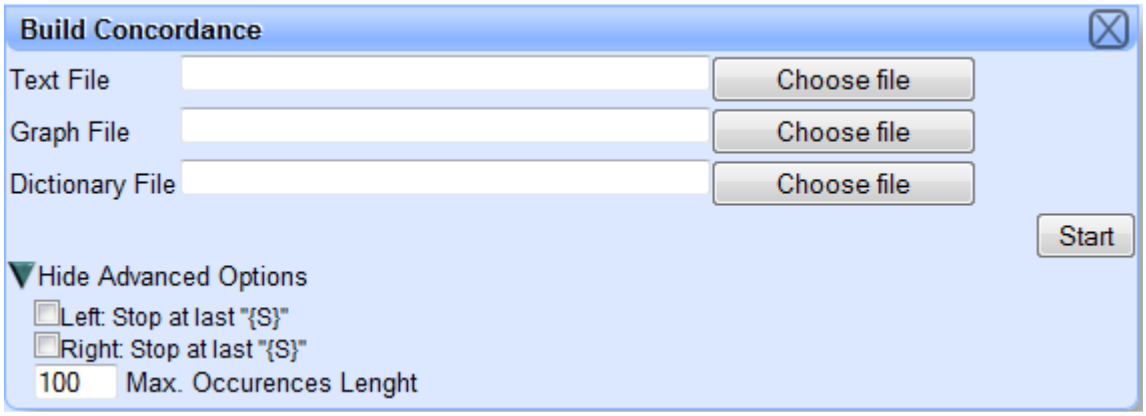
a) Open Applet : here you can call the Computerlingustic Tool UNITEX, which is developed as a public domain tool at: Université Paris-Est Marne-la-Vallée, see: http://www-igm.univ-mlv.fr/~unitex/ The tool is opened in a Separate TAB: “CisWeb – unitex” of your Browser:


CisWeb - unitex - Mozilla Firefox
Datei Bearbeiten Ansicht Chronik Lesezeichen Extras Hilfe
CisWeb - unitex
CisWeb Workbench
http://cisweb64.cis.uni-muenchen.de:8090/CisWeb/CisWeb.html
Unitex
Text DELA FSGraph Lexicon-Grammar Edit File Edition Windows Info

Picture 8:UNITEX as JAVA-Applet

b) Run Script: here you can run UNITEX with a PERL Script via the command line.

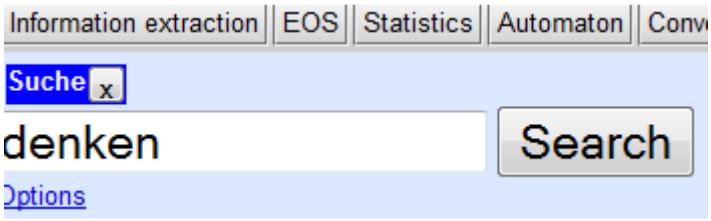
c) Concordance: Here a new Concordance Tool is under Development, with allows to locate pattern in a text file, which can be specified as local grammar. The tools accepts a Local Grammar as Graph-File and a Electronic-Dictionary in DELA-Format. With the help of the Dictionary the tool can work lemmatized on the text. The tool uses an implementation of automaton, developed by Florian Fink (CIS, 2011). The Concordance is programmed from Patrick Seebauer (CIS, 2010):



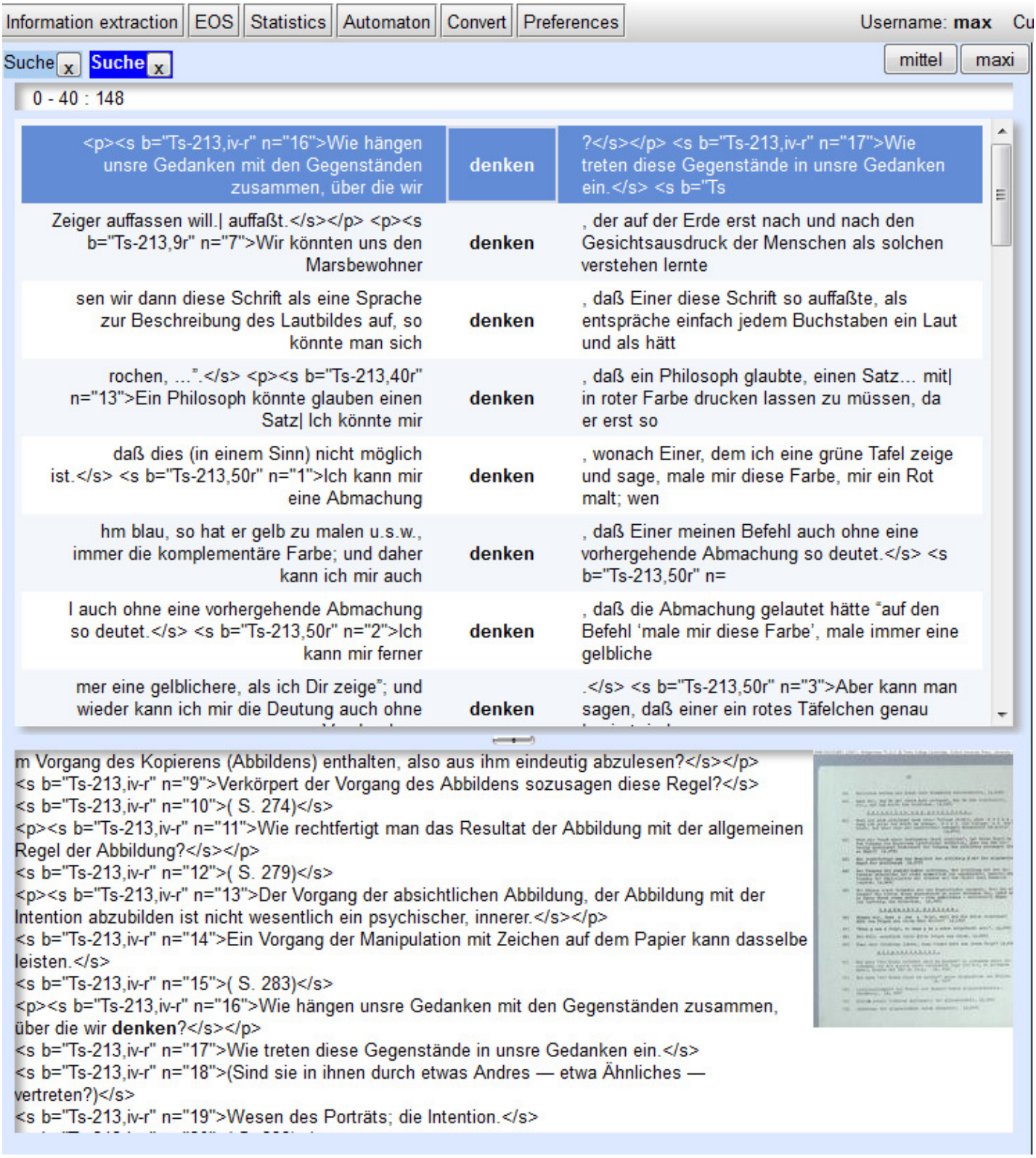
Picture 9: Concordance Search and Concordance Viewer

d) Suche -> Wittgenstein: This tool is connected to the Big Typescript of Ludwig Wittgenstein, http://www.wittgensteinsource.org/ : "Wittgenstein-Typescript 213 mit freundlicher Genehmigung von: Wittgenstein Source, Bergen Text Edition (BTE), wittgensteinsource.org
Edited by Alois Pichler, in collaboration with H.W. Krüger, D.C.P. Smith, T.M. Bruvik, V. Olstad; Wittgenstein Archives at the University of Bergen, Uni Digital, Research Cooperation: Alois Pichler and Max Hadersbeck"

If your CisWeb Account has the Right to read the TS-213 Big Typescript, the tool locates the Pattern, which is supplied in the Menu. The tool uses an implementation of automaton, developed by Florian Fink (CIS, 2011). The Concordance is programmed from Patrick Seebauer (CIS, 2010):

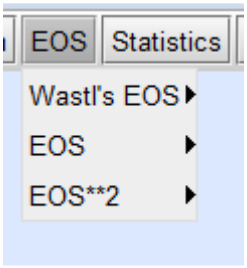


Picture 9: Concordance Search and Concordance Viewer connected to the Big Typescript of Ludwig Wittgenstein.



Picture 10: Output of the Concordance and Search Tool

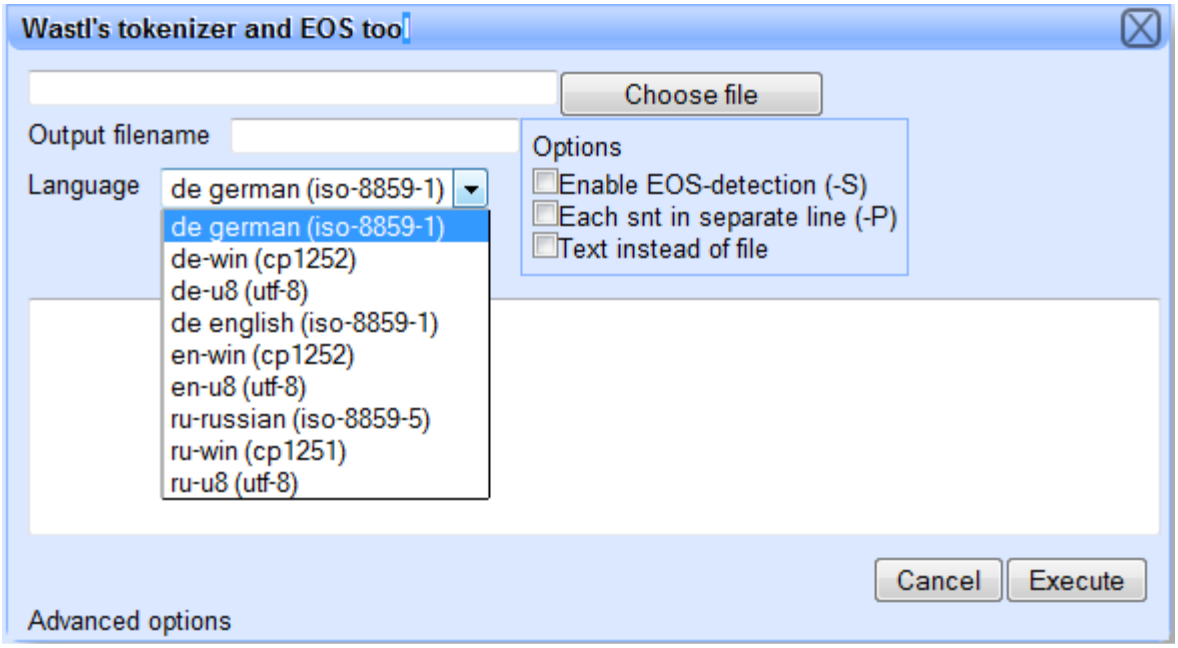
5.4 EOS (End-Of-Sentence recognizer)



Picture 11: End of Sentence Menu

Behind the EOS-Menu 3 different End-Of-Sentence Tools are hidden. A textfile (Choose File) is parsed and at the end of a sentence an EOS-mark is added. The Result is stored in an output file (Output filename) in your working directory.
If you don't specify an output Filename, a resultfile with the name "$(Input filename)_eos" will be added to your working directory.

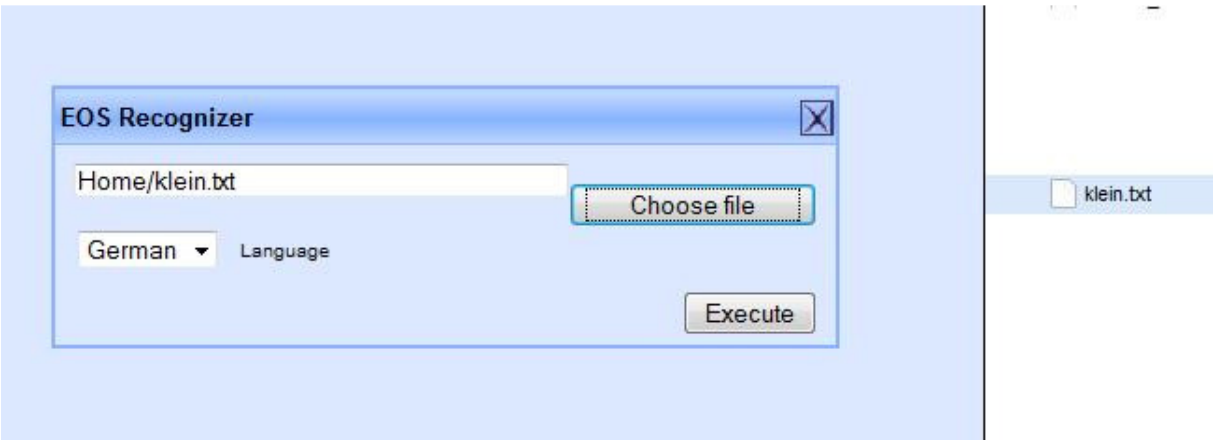
a) Wastl's EOS: This tool is developed from Sebastian Nagel, a former researcher at the CIS. See: http://www.cis.uni-muenchen.de/~wastl/



Picture 12: End of Sentence detection from Sebastian Nagel

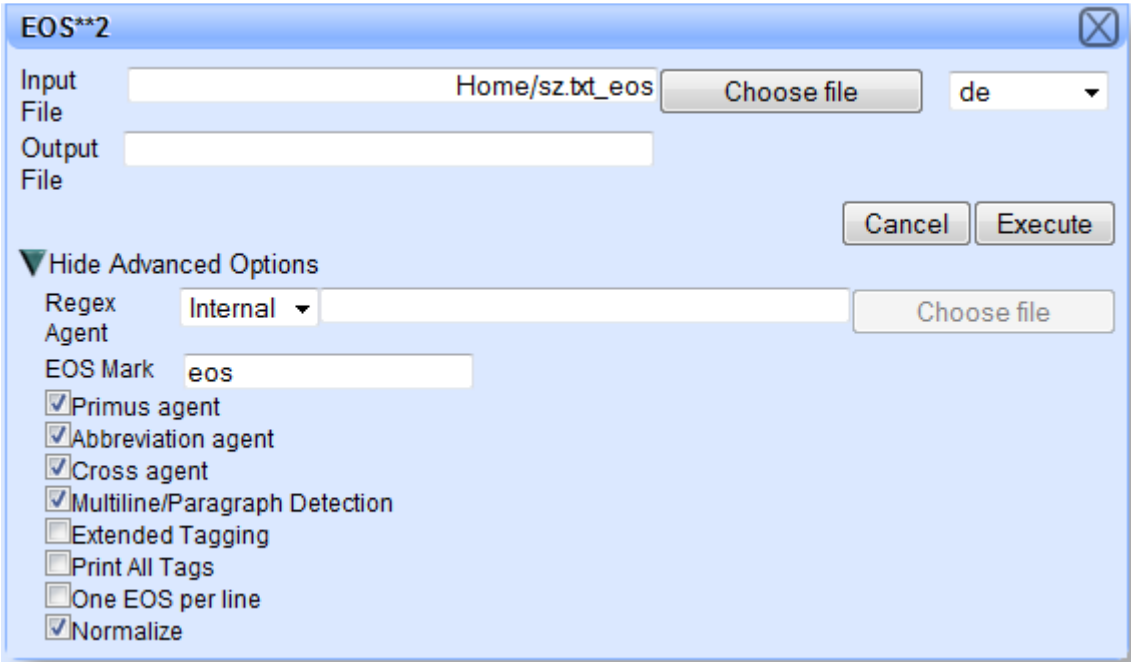
This tool allows to Segment German, Englisch and even Russian Textfiles stored in different encodings.

b) EOS: This is the first version of an speed-optimized C-Version : EndOfSentence, See: Demopage: http://maxdemo.cis.uni-muenchen.de/home_demos/eosv1/index.html



Picture 13: End of Sentence detection First C-Version

c) EOS++: This is the actual C++ Version of EndOfSentence Detection. This version is permanently enhanced and new languages are added: (French, Italian, Croatian, Russian).
See: Demopage: http://maxdemo.cis.uni-muenchen.de/home_demos/eos/index.html
See: Readme-File: http://maxdemo.cis.uni-muenchen.de/home_demos/eos/readme.htm



Picture 14: End of Sentence detection actual C++ Version

5.6. Statistics

- Language:

Detection of a language of a textfile. Caution: Files with UTF16 Encoding are not supported.

- Frequency List

Produces a Frequency List of an utf8 textfile.

5.7. Automaton

- TGC Gramatics

Tgc (Transducer Grammar Compiler) compiles Transducer-Grammars to openfst-Transducers, implemented from Jörg Schuster.

5.8. Convert

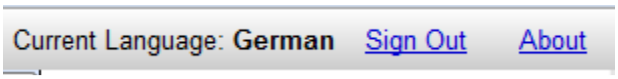
- Recode

recodes a File from a selectable Encoding to another Encoding.

- Strip HTML

Stripps all HTML-Tags within an .html formatted File and produces a Textfile out of it.

6. Logout, Info's and Wiki's of CisWeb



Picture 15: CisWeb logout, About

In the upper right corner of CisWeb you will find the "logout" Option of CisWeb. In this corner you also find an Option "About", were you can find an Link to the CisWeb wiki, where you get additional Documentation about CisWeb, can deliver wishes for new features, which should be implemented in CisWeb. In the Wiki you also find a WiKi Page, where you can report error's, which happen in CisWeb.

7. Prospect

The current version is only a picture of the state of the art, a lot of useful functionalities are not yet implemented. However the range of use of this application is big. One could think of using it as a lecture attendant tool where the professor/tutor can make his text files available in the public folder. The students could work with the tools, provided by this web service without having to use a shell to execute the tools but a graphical user interface. Furthermore every command line tool can be integrated into CisWeb very easily. You only have to add a GUI at the client side and implement the RPC. And it is a good chance to present CIS to the rest of the world because everybody who has a browser and internet can access the web service. Even from Singapore it worked very well! There are many possibilities how to upgrade this application.
But that's dreams of the future...